Recall: Linear Regression

\[ f(x_i) = \mathbf{w}^T \mathbf{x}_i \]
Recall: Linear Regression

\[ y_i = \overline{w}^T \overline{x}_i + \epsilon_i \text{ where all } \epsilon_i \text{ are i.i.d. } N(0, \sigma^2) \]
Before we see any training data...

... we must believe that all hypotheses are possible and that all features behave identically.

However, by Occam’s Razor, we might believe that hypotheses with smaller weights are more likely

\[ w_d \sim N(0,1) \quad \forall \, d \in \{0, \ldots, D\} \]

\[ \vec{w} \sim N(\vec{0}_{D+1}, I_{D+1}) \] where \( \vec{0}_{D+1} \) is a vector of \( D + 1 \) 0’s and \( I_{D+1} \) is the \( (D + 1) \)-by-\( (D + 1) \) identity matrix
Intuition: given some training data, find the most likely weight vector

Tradeoff between two (generally) competing factors:
- The prior on $\overrightarrow{w}$ makes small weight vectors more likely
- The prior on the residuals $\epsilon_i$ makes weight vectors with low training error more likely
After we see training data, we have to update our prior belief to reflect what we saw.

Given $\mathcal{D} = \left\{ X = [\overrightarrow{x_1}, \ldots, \overrightarrow{x_n}] \in \mathbb{R}^{D+1 \times n}, \overrightarrow{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n \right\}$

Find $\overrightarrow{w}^* = \arg \max P\{\overrightarrow{w} | X, \overrightarrow{y}\}$ using Bayes rule:

$$P\{\overrightarrow{w} | X, \overrightarrow{y}\} = \frac{P\{\overrightarrow{w}, \overrightarrow{y} | X\}}{P\{\overrightarrow{y} | X\}}$$
A Posterior Belief

- After we see training data, we have to update our prior belief to reflect what we saw.

- Given \( \mathcal{D} = \left\{ X = [x_1 \ldots x_n] \in \mathbb{R}^{D+1 \times n}, \hat{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n \right\} \)

- Find \( \bar{w}^* = \arg\max P\{\bar{w} | X, \hat{y}\} \) using Bayes rule:

\[
P\{\bar{w} | X, \hat{y}\} = \frac{P\{\bar{w}, \hat{y} | X\}}{P\{\hat{y} | X\}} = \frac{P\{\hat{y} | X, \bar{w}\}P\{\bar{w}\}}{P\{\hat{y} | X\}}
\]
Definitions

\[ P(\vec{w} \mid X, \hat{y}) = \frac{P(\vec{w}, \hat{y} \mid X)}{P(\hat{y} \mid X)} = \frac{P(\hat{y} \mid X, \vec{w})P(\vec{w})}{P(\hat{y} \mid X)} \]

- \( P(\vec{w} \mid X, \hat{y}) \) is called the posterior
- \( P(\hat{y} \mid X, \vec{w}) \) is called the likelihood i.e. how likely are the labels given the locations and the weights
- \( P(\vec{w}) \) is called the prior
- \( P(\hat{y} \mid X) \) is called the marginal likelihood i.e. it’s the likelihood with the weight vector “marginalized” out
  - \( P(\hat{y} \mid X) = \int P(\hat{y} \mid X, \vec{w})P(\vec{w})d\vec{w} \)
\[
P\{\tilde{y} \mid X, \tilde{w}\} = P\{\epsilon_1 = y_1 - \tilde{w}^T \tilde{x}_1 \cap \ldots \cap \epsilon_n = y_n - \tilde{w}^T \tilde{x}_n\}
\]
\[
= \prod_{i=1}^{n} P\{\epsilon_i = y_i - \tilde{w}^T \tilde{x}_i\} \quad (\epsilon_i \sim N(0, \sigma^2))
\]
\[
= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \tilde{w}^T \tilde{x}_i - 0)^2\right)
\]
\[
= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \tilde{w}^T \tilde{x}_i)^2\right)
\]
\[
= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{(\tilde{y} - \tilde{w}^T X)^T (\tilde{y} - \tilde{w}^T X)}{2\sigma^2}\right)
\]
\[
\tilde{y} \mid X, \tilde{w} \sim N(\tilde{w}^T X, \sigma^2 I_n)
\]
\[ P\{ \vec{w} | X, \vec{y} \} = \frac{P\{ \vec{y} | X, \vec{w} \} P\{ \vec{w} \}}{P\{ \vec{y} | X \}} = \frac{P\{ \vec{y} | X, \vec{w} \} P\{ \vec{w} \}}{\int P\{ \vec{y} | X, \vec{w} \} P\{ \vec{w} \} d\vec{w}} \]

- \( \vec{w} \sim N(\vec{0}_{D+1}, I_{D+1}) \)
- \( \vec{y} | X, \vec{w} \sim N(\vec{w}^T X, \sigma^2 I_n) \)
- \( \vec{w} | X, \vec{y} \sim N(\vec{\mu}, \Sigma^{-1}) \)
  - \( \Sigma = \frac{XX^T}{\sigma^2} + I_{D+1} \)
  - \( \vec{\mu} = \Sigma^{-1} \frac{(X\vec{y})}{\sigma^2} \)
  - \( \vec{w}^* = \vec{\mu} \)
- \( g(\vec{x}) = \vec{w}^* \vec{x} = \vec{\mu}^T \vec{x} \)
Given the distribution of $\mathbf{w} \mid X, \mathbf{y}$, we can reason about the distribution of $g(\mathbf{x}) \mid X, \mathbf{y} = \mathbf{w}^T \mathbf{x} \mid X, \mathbf{y}$

- If $\mathbf{w} \mid X, \mathbf{y} \sim N(\bar{\mu}, \Sigma^{-1})$
- ... then $\mathbf{w}^T \mathbf{x} \mid X, \mathbf{y} \sim N(\bar{\mu}^T \bar{x}, \bar{x}^T \Sigma^{-1} \bar{x})$

Instead of predicting a single value, we can now make probabilistic predictions for query points
Bayesian Linear Regression

- Given the distribution of $\vec{w} \mid X, \vec{y}$, we can reason about the distribution of $g(\vec{x}) \mid X, \vec{y} = \vec{w}^T \vec{x} \mid X, \vec{y}$

- If $\vec{w} \mid X, \vec{y} \sim N(\vec{\mu}, \Sigma^{-1})$
- ... then $\vec{w}^T \vec{x} \mid X, \vec{y} \sim N(\vec{\mu}^T \vec{x}, \vec{x}^T \Sigma^{-1} \vec{x})$

- Instead of predicting a single value, we can now make probabilistic predictions for query points
Bayesian Linear Regression

- Given $\mathcal{D} = \{X, \tilde{y}\}, g(\tilde{x}) | X, \tilde{y} \sim N(\tilde{\mu}^T \tilde{x}, \tilde{x}^T \Sigma^{-1} \tilde{x})$
  
where $\Sigma = \frac{XX^T}{\sigma^2} + I_{D+1}$ and $\tilde{\mu} = \Sigma^{-1} \frac{(X\tilde{y})}{\sigma^2}$

- Bayesian linear regression is an inner product method!
• Find a maximal-margin separating hyperplane in the transformed space, $[\overline{w}^*, w_0^*]$, by solving the QP:

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2} \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \\
\text{subject to} & \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \quad \alpha_i \geq 0 \quad \forall \ i \in \{1, \ldots, n\}
\end{align*}$$

• Return the corresponding predictor in the original space:

$$g(\overline{x}) = \text{sign}\left( \sum_{i: \alpha_i^* > 0} \alpha_i^* y_i \overline{x}_i^T \overline{x}_j + w_0^* \right)$$
Bayesian Linear Regression

• Given $D = \{X, \tilde{y}\}$, $g(\tilde{x}) | X, \tilde{y} \sim N(\tilde{\mu}^T \tilde{x}, \tilde{x}^T \Sigma^{-1} \tilde{x})$

  where $\Sigma = \frac{XX^T}{\sigma^2} + I_{D+1}$ and $\tilde{\mu} = \Sigma^{-1} \frac{(X\tilde{y})}{\sigma^2}$

• Bayesian linear regression is an inner product method!

• $\tilde{\mu}^T \tilde{x} = \tilde{x}^T X (X^T X + \sigma^2 I_n)^{-1} \tilde{y}$

• $\tilde{x}^T \Sigma^{-1} \tilde{x} = \tilde{x}^T \tilde{x} - \tilde{x}^T X (X^T X + \sigma^2 I_n)^{-1} (\tilde{x}^T X)^T$
Bayesian Linear Regression

- Given $\mathcal{D} = \{X, y\}$, $g(\tilde{x}) | X, y \sim N(\tilde{\mu}^T \tilde{x}, \tilde{x}^T \Sigma^{-1} \tilde{x})$

  \[
  \Sigma = \frac{X I_n X^T}{\sigma^2} + I_{D+1} \text{ and } \tilde{\mu} = \Sigma^{-1} (X y)
  \]

- Bayesian linear regression is an inner product method!

- $\tilde{\mu}^T \tilde{x} = \tilde{x}^T X (X^T I_{D+1} X + \sigma^2 I_n)^{-1} y$

- $\tilde{x}^T \Sigma^{-1} \tilde{x} = \tilde{x}^T \tilde{x} - \tilde{x}^T X (X^T I_{D+1} X + \sigma^2 I_n)^{-1} (\tilde{x}^T X)^T$
Bayesian Linear Regression

- Given \( D = \{X, \hat{y}\}, g(\hat{x}) | X, \hat{y} \sim N(\mu^T \hat{x}, \hat{x}^T \Sigma^{-1} \hat{x}) \)

  where \( \Sigma = \frac{XX^T}{\sigma^2} + I_D + 1 \) and \( \mu = \Sigma^{-1} \left(\frac{X \hat{y}}{\sigma^2}\right) \)

- Bayesian linear regression is an inner product method!

  \[ \hat{\mu}^T \hat{x} = \hat{x}^T X (X^T X + \sigma^2 I_n)^{-1} \hat{y} \]

  \[ \hat{x}^T \Sigma^{-1} \hat{x} = \hat{x}^T \hat{x} - \hat{x}^T X (X^T X + \sigma^2 I_n)^{-1} (\hat{x}^T X)^T \]

  \[ X^T X = \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \cdots & x_1^T x_n \\ x_2^T x_1 & x_2^T x_2 & \cdots & x_2^T x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n^T x_1 & x_n^T x_2 & \cdots & x_n^T x_n \end{bmatrix} \]
• Find a maximal-margin separating hyperplane in the transformed space, $[\tilde{w}^*, \tilde{w}_0^*]$, by solving the QP:

$$\minimize \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{\alpha}_i \tilde{\alpha}_j y_i y_j \Phi(x_i^*)^T \Phi(x_j^*) - \sum_{i=1}^{n} \tilde{\alpha}_i$$

subject to $\sum_{i=1}^{n} \tilde{\alpha}_i y_i = 0$

$\tilde{\alpha}_i \geq 0 \ \forall \ i \in \{1, ..., n\}$

• Return the corresponding predictor in the original space:

$$g(\tilde{x}) = \text{sign} \left( \sum_{i: \tilde{\alpha}_i^* > 0} \tilde{\alpha}_i^* y_i \Phi(x_i^*)^T \Phi(\tilde{x}) + \tilde{w}_0^* \right)$$
• Find a maximal-margin separating hyperplane in the transformed space, $[\tilde{w}^*, \tilde{w}_0^*]$, by solving the QP:

$$\min \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{\alpha}_i \tilde{\alpha}_j y_i y_j K(\tilde{x}_i, \tilde{x}_j) - \sum_{i=1}^{n} \tilde{\alpha}_i$$

subject to

$$\sum_{i=1}^{n} \tilde{\alpha}_i y_i = 0$$

$$\tilde{\alpha}_i \geq 0 \forall i \in \{1, ..., n\}$$

• Return the corresponding predictor in the original space:

$$g(\tilde{x}) = \text{sign} \left( \sum_{i: \tilde{\alpha}_i > 0} \tilde{\alpha}_i^* y_i K(\tilde{x}_i, \tilde{x}_j) + \tilde{w}_0^* \right)$$
Bayesian Linear Regression

• Given \( D = \{X, \tilde{y}\}, \ g(\tilde{x}) \mid X, \tilde{y} \sim N(\tilde{\mu}^T\tilde{x}, \tilde{x}^T\Sigma^{-1}\tilde{x}) \)

where \( \Sigma = \frac{XX^T}{\sigma^2} + I_{D+1} \) and \( \tilde{\mu} = \Sigma^{-1}\frac{(XY)}{\sigma^2} \)

• \( \tilde{\mu}^T\tilde{x} = \tilde{x}^T X(X^TX + \sigma^2 I_n)^{-1}\tilde{y} \)

• \( \tilde{x}^T\Sigma^{-1}\tilde{x} = \tilde{x}^T \tilde{x} - \tilde{x}^T X(X^TX + \sigma^2 I_n)^{-1}(\tilde{x}^T X)^T \)
Given $\mathcal{D} = \{X, \tilde{y}\}$ and a nonlinear transformation $\Phi$, $g(\tilde{x}) | X, \tilde{y} \sim N(\hat{\mu}^T \Phi(\tilde{x}), \Phi(\tilde{x})^T \Sigma^{-1} \Phi(\tilde{x}))$

where $\Sigma = \frac{\Phi(x)\Phi(x)^T}{\sigma^2} + I_{D+1}$ and $\hat{\mu} = \Sigma^{-1} \frac{\Phi(x)\tilde{y}}{\sigma^2}$

- $\hat{\mu}^T \Phi(\tilde{x}) = \Phi(\tilde{x})^T \Phi(X)(\Phi(X)^T \Phi(X) + \sigma^2 I_n)^{-1} \tilde{y}$,

- $\Phi(\tilde{x})^T \Sigma^{-1} \Phi(\tilde{x}) = \Phi(\tilde{x})^T \Phi(\tilde{x}) - $ $\Phi(\tilde{x})^T \Phi(X)(\Phi(X)^T \Phi(X) + \sigma^2 I_n)^{-1}(\Phi(\tilde{x})^T \Phi(X))^T$

- Let $K(\tilde{x}, \tilde{x}') = \Phi(\tilde{x})^T \Phi(\tilde{x}')$
Given $\mathcal{D} = \{X, \tilde{y}\}$ and a nonlinear transformation $\Phi$, $g(\tilde{x}) | X, \tilde{y} \sim N(\tilde{\mu}^T \Phi(\tilde{x}), \Phi(\tilde{x})^T \Sigma^{-1} \Phi(\tilde{x}))$

where $\Sigma = \frac{\Phi(x)\Phi(x)^T}{\sigma^2} + I_{D+1}$ and $\tilde{\mu} = \Sigma^{-1} \frac{(\Phi(x)\tilde{y})}{\sigma^2}$

- $\tilde{\mu}^T \Phi(\tilde{x}) = K(\tilde{x}, X)(K(X, X) + \sigma^2 I_n)^{-1} \tilde{y}$,
- $\Phi(\tilde{x})^T \Sigma^{-1} \Phi(\tilde{x}) = K(\tilde{x}, \tilde{x}) - K(\tilde{x}, X)(K(X, X) + \sigma^2 I_n)^{-1} (K(\tilde{x}, X))^T$
- Let $K(\tilde{x}, \tilde{x}') = \Phi(\tilde{x})^T \Phi(\tilde{x}')$
Gaussian Process Regression

- Given $\mathcal{D} = \{X, \tilde{y}\}$ and a kernel $K$,

$$g(\tilde{x}) \mid X, \tilde{y} \sim N\left(K(\tilde{x}, X)(K(X, X) + \sigma^2 I_n)^{-1} \tilde{y}, K(\tilde{x}, \tilde{x}) - K(\tilde{x}, X)(K(X, X) + \sigma^2 I_n)^{-1}(K(\tilde{x}, X))^T\right)$$
• The target function $f$ is unknown

• Place a probability distribution directly on $f$

• A Gaussian process (GP) is like a Gaussian distribution over functions
  • Imagine a function as a collection of infinitely many random variables, one at each location in the input
  • A Gaussian process can be thought of as an infinite-dimensional multivariate Gaussian distribution
<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Variable</th>
<th>Mean</th>
<th>(Co)Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate</td>
<td>$x$</td>
<td>A scalar, $\mu$</td>
<td>A scalar, $\sigma^2$</td>
</tr>
<tr>
<td>Variable</td>
<td>Mean</td>
<td>(Co)Variance</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>$x$</td>
<td>A scalar, $\mu$</td>
<td>A scalar, $\sigma^2$</td>
</tr>
<tr>
<td>Multivariate</td>
<td>$\tilde{x}$</td>
<td>A vector, $\tilde{\mu}$</td>
<td>A matrix, $\Sigma$</td>
</tr>
</tbody>
</table>

### Gaussians
Gaussian Processes

- Extension of the multivariate Gaussian distribution from a finite set of random variables to a function

- Fully defined by a mean function, $\mu(\cdot)$, and a covariance function, $K(\cdot, \cdot)$
  - The mean function can be describes the central tendency of the function
  - The covariance function describes something about the structure/shape of the function.
A GP belief on $f$ induces a multivariate Gaussian belief on the value of $f$ at any finite set of points.

If $X = [\overrightarrow{x_1} \cdots \overrightarrow{x_n}]$ and $f \sim GP(\mu, K)$ ...

... then $f(X) \sim N(\mu(X), K(X,X))$

\[
\mu(X) = [\mu(\overrightarrow{x_1}) \cdots \mu(\overrightarrow{x_n})]
\]

\[
K(X,X) = \begin{bmatrix}
K(\overrightarrow{x_1}, \overrightarrow{x_1}) & K(\overrightarrow{x_1}, \overrightarrow{x_2}) & \cdots & K(\overrightarrow{x_1}, \overrightarrow{x_n}) \\
K(\overrightarrow{x_2}, \overrightarrow{x_1}) & K(\overrightarrow{x_2}, \overrightarrow{x_2}) & \cdots & K(\overrightarrow{x_2}, \overrightarrow{x_n}) \\
\vdots & \vdots & \ddots & \vdots \\
K(\overrightarrow{x_n}, \overrightarrow{x_1}) & K(\overrightarrow{x_n}, \overrightarrow{x_2}) & \cdots & K(\overrightarrow{x_n}, \overrightarrow{x_n})
\end{bmatrix}
\]
• GPs are closed under conditioning
  • If you have a GP belief on $f$ and you have some training data $\mathcal{D} = \{X, \hat{y}\}$ ...
  • ... then you can condition your prior GP belief on $f$ using $\mathcal{D}$ to get a posterior belief that is still a GP!
  • Specifically, if $f \sim GP(\mu, K)$ then $f \mid \mathcal{D} \sim GP(\mu_\mathcal{D}, K_\mathcal{D})$
  • $\mu_\mathcal{D}(\tilde{x}) = \mu(\tilde{x}) + K(\tilde{x}, X)(K(X, X) + \sigma^2 I_n)^{-1}(\hat{y} - \mu(X))$
  • $K_\mathcal{D}(\tilde{x}, \tilde{x}) = K(\tilde{x}, \tilde{x}) - K(\tilde{x}, X)(K(X, X) + \sigma^2 I_n)^{-1}(K(X, \tilde{x}))^T$
GP Prior

\[ \pm 2\sqrt{K(x,x)} \]

\[ \mu(x) \]
GP Samples
GP Posterior

RBF Kernel: $K(\tilde{x}, \tilde{x}') = e^{-\frac{||\tilde{x} - \tilde{x}'||^2}{2}}$
GP Posterior

RBF Kernel: $K(\mathbf{x}, \mathbf{x}') = \exp \left( \frac{-||\mathbf{x} - \mathbf{x}'||^2}{2} \right)$
GP
Posterior

\[ K(\hat{x}, \hat{x}') = ??? \]
GP Posterior

\[ K(\mathbf{x}, \mathbf{x'}) = \min(\mathbf{x}, \mathbf{x'}) \]
GP Pros and Cons

• Pros:
  • Probabilistic predictions allow for uncertainty quantification
  • Can encode prior knowledge via the covariance function/kernel

• Cons:
  • Inference requires an $n$-by-$n$ matrix inversion
  • Can encode prior misconceptions via the covariance function/kernel...